

SMML estimators for linear regression and tessellations of hyperbolic space

James G. Dowty

March 19, 2014

Abstract

The strict minimum message length (SMML) principle links data compression with inductive inference. The corresponding estimators have many useful properties but they can be hard to calculate. We investigate SMML estimators for linear regression models and we show that they have close connections to hyperbolic geometry. When equipped with the Fisher information metric, the linear regression model with p covariates and a sample size of n becomes a Riemannian manifold, and we show that this is isometric to $(p+1)$ -dimensional hyperbolic space \mathbb{H}^{p+1} equipped with a metric tensor which is $2n$ times the usual metric tensor on \mathbb{H}^{p+1} . A natural identification then allows us to also view the set of sufficient statistics for the linear regression model as a hyperbolic space. We show that the partition of an SMML estimator corresponds to a tessellation of this hyperbolic space.

1 The linear regression model

To establish our notation we briefly recall some details of the linear regression model.

The linear regression model is a statistical model for observed data $y \in \mathbb{R}^n$ (thought of as a column matrix) which is a realization of an n -dimensional, normally-distributed random variable Y with mean $A\beta$ and variance-covariance matrix $\sigma^2 I_n$, i.e.,

$$Y \sim N_n(A\beta, \sigma^2 I_n),$$

where A is a full-rank $n \times p$ matrix called the design matrix, $\beta \in \mathbb{R}^p$ is a column matrix, $\sigma > 0$ and I_n is the $n \times n$ identity matrix. Here β and σ are unknown and are to be estimated in terms of y and A . In this paper, we will always require $p \leq n$ though for certain results (indicated in the text) we will also require $p < n$. The probability density function (PDF) of Y given values of the unknown model parameters β and σ is therefore

$$(2\pi\sigma^2)^{-n/2} \exp\left(-\frac{\|y - A\beta\|^2}{2\sigma^2}\right) \quad (1)$$

where $\|\cdot\|$ is the Euclidean norm on \mathbb{R}^n .

It is well-known that this statistical model is an exponential family, so we will now write (1) in canonical form. Let B be any $n \times p$ matrix whose columns form an orthonormal basis for the column space $\text{col } A$ of A , e.g. we could take $B = A(A^T A)^{-\frac{1}{2}}$. Then $B^T B = I_p$ and

the orthogonal projection of \mathbb{R}^n onto $\text{col } A$ is $A(A^T A)^{-1} A^T = BB^T$. Define the *sufficient statistics* $T(y)$ and *natural parameters* θ of the exponential family to be

$$T(y) \stackrel{\text{def}}{=} \begin{bmatrix} B^T y \\ \|y\|^2 \end{bmatrix} \quad \text{and} \quad \theta \stackrel{\text{def}}{=} \frac{1}{\sigma^2} \begin{bmatrix} B^T A \beta \\ -\frac{1}{2} \end{bmatrix}. \quad (2)$$

Then the PDF (1) can be written in the canonical form

$$p_Y(y|\theta) = \exp(\theta \cdot T(y)) h_Y(y) / Z(\theta) \quad (3)$$

where the dot denotes the Euclidean inner product, $h_Y(y) = (2\pi)^{-n/2}$ and the *partition function* $Z(\theta)$ is

$$Z(\theta) \stackrel{\text{def}}{=} \exp \left(-\frac{n}{2} \log(-2\theta_{p+1}) - \frac{\theta_1^2 + \dots + \theta_p^2}{4\theta_{p+1}} \right). \quad (4)$$

Note from (2) that the natural parameter space Θ , which is the set of all natural parameters, is

$$\Theta = \{\theta \in \mathbb{R}^{p+1} \mid \theta_{p+1} < 0\}. \quad (5)$$

Remark 1. The first p sufficient statistics $B^T y$ are essentially equal to the orthogonal projection of y onto $\text{col } A$. More precisely, since $B^T y = B^T (BB^T y)$ and BB^T is orthogonal projection, the first p sufficient statistics are the orthogonal projection of y onto $\text{col } A$ written in terms of the co-ordinates for $\text{col } A$ corresponding to the basis formed by the columns of B . The reason for using this definition, instead of simply taking the orthogonal projection of y onto $\text{col } A$, is that we require the set of all possible sufficient statistics to form an open set in \mathbb{R}^d for some d , while $\text{col } A$ is a lower-dimensional set in \mathbb{R}^n .

Remark 2. In this paper, we will think of Θ as simply being a subset of a generic $(p+1)$ -dimensional vector space \mathbb{R}^{p+1} . However, for a number of reasons, it is more natural to think of Θ as a subset of the dual space to the vector space containing the set \mathcal{X} of all possible sufficient statistics. One reason this is natural is that the dot in (3) then becomes the natural pairing between a vector space and its dual, rather than the (non-canonical) Euclidean dot product. Another reason is that the Fisher information matrices on Θ and \mathcal{X} (when \mathcal{X} is identified with the expectation parameter space, see Section 4.2) are matrix inverses of each other, as is the case for a metric on a vector space and the induced metric on the dual vector space. A third reason is that there is a close connection between exponential families and convex conjugation [1, Chapter 9] which makes it natural to think of Θ and \mathcal{X} as convex subsets of dual vector spaces. This connection can be used to show (under mild conditions) that the maximum likelihood estimator is the gradient of the maximized log-likelihood function, and that the maximized log-likelihood function can itself be calculated as the convex conjugate of the log-partition function [1, Theorem 9.13].

2 The linear regression model is isometric to $2n\mathbb{H}^{p+1}$

When equipped with the Fisher information metric, the parameter space for the linear regression model above, with p covariates and a sample size of n , is a Riemannian manifold. In this section, we will show that this is isometric to the Riemannian manifold $2n\mathbb{H}^{p+1}$, which we define to be $(p+1)$ -dimensional hyperbolic space \mathbb{H}^{p+1} (with all sectional curvatures equal to -1) equipped with a metric tensor which is $2n$ times the usual metric tensor on \mathbb{H}^{p+1} . This result contradicts certain findings of [4] and [2] when $n \neq 1$, but we will show that the formulae of [4] and [2] are not correct.

Recall that if an open set $U \subseteq \mathbb{R}^k$ parameterises a stochastic model then the *Fisher information metric* of this model is represented, in the local co-ordinates of this parameterisation, by the *Fisher information matrix* g_U . Under regularity conditions satisfied by all models considered in this paper, g_U is given by either of the following expressions

$$g_U = \mathbb{E}[(\nabla \ell)(\nabla \ell)^T] = -\mathbb{E}[\text{Hess}(\ell)] \quad (6)$$

where $\ell : U \rightarrow \mathbb{R}$ is the log-likelihood function, $\nabla \ell$ is its gradient (interpreted as a column matrix in the formula above), $\text{Hess}(\ell)$ is its Hessian matrix and the expectation is taken over the observed data.

2.1 The upper half-space parameterisation

We now define a parameterisation for the linear regression model and calculate its corresponding Fisher information matrix. Let

$$\phi \stackrel{\text{def}}{=} \begin{bmatrix} B^T A \beta \\ \sigma \sqrt{2n} \end{bmatrix} \quad (7)$$

and note that, up to a linear transformation, this is just the β, σ parameterisation. The set of possible values for ϕ is the upper half-space $\Phi \stackrel{\text{def}}{=} \{\phi \in \mathbb{R}^{p+1} \mid \phi_{p+1} > 0\}$ so, in light of this and Theorem 1 below, we will refer to this as the *upper half-space parameterisation*.

The upper half-space model for hyperbolic space is a Riemannian manifold with a metric tensor which is a particular multiple of the identity, as given in [5, Theorem 4.6.6], and all sectional curvatures equal to -1 .

Theorem 1. *The Fisher information matrix for the upper half-space parameterisation is*

$$g_\Phi = 2n\phi_{p+1}^{-2} I_{p+1}$$

where I_{p+1} is the $(p+1) \times (p+1)$ identity matrix. So Φ is the upper half-space model for $(p+1)$ -dimensional hyperbolic space but with a metric tensor that is $2n$ times the usual metric tensor.

Proof. We first note that $A\beta = B\phi_{[1:p]}$, where $\phi_{[1:p]} = B^T A\beta$ is the $p \times 1$ column matrix whose entries are the first p entries of ϕ . This follows because BB^T is the identity on $\text{col } A$ (being the orthogonal projection onto $\text{col } A$) and $A\beta \in \text{col } A$ so $A\beta = BB^T A\beta = B\phi_{[1:p]}$. So by (1), the log-likelihood function for this parameterisation is

$$\ell_\Phi(\phi) = -\frac{n}{2} \log(\pi/n) - n \log \phi_{p+1} - n\phi_{p+1}^{-2} \|y - B\phi_{[1:p]}\|^2.$$

For $i, j = 1, \dots, p$ we therefore have

$$\frac{\partial \ell_\Phi}{\partial \phi_i} = 2n\phi_{p+1}^{-2} (y - B\phi_{[1:p]})^T B e_i,$$

where e_i is the i^{th} standard basis vector for \mathbb{R}^p , and

$$\frac{\partial \ell_\Phi}{\partial \phi_{p+1}} = -n\phi_{p+1}^{-1} + 2n\phi_{p+1}^{-3} \|y - B\phi_{[1:p]}\|^2.$$

So letting δ_{ij} be the Kronecker delta,

$$\frac{\partial^2 \ell_\Phi}{\partial \phi_i \partial \phi_j} = -2n\phi_{p+1}^{-2} \delta_{ij},$$

$$\frac{\partial^2 \ell_\Phi}{\partial \phi_i \partial \phi_{p+1}} = -4n\phi_{p+1}^{-3} (y - B\phi_{[1:p]})^T B e_i$$

and

$$\frac{\partial^2 \ell_\Phi}{\partial \phi_{p+1}^2} = n\phi_{p+1}^{-2} - 6n\phi_{p+1}^{-4} \|y - B\phi_{[1:p]}\|^2.$$

Taking expectations of the negatives of these second partial derivatives and using the facts $\mathbb{E}[y] = A\beta = B\phi_{[1:p]}$ and

$$\mathbb{E}\|y - B\phi_{[1:p]}\|^2 = \sum_{i=1}^n \mathbb{E}[(y_i - \mathbb{E}[y_i])^2] = n\sigma^2 = \frac{\phi_{p+1}^2}{2}$$

then proves $g_\Phi = 2n\phi_{p+1}^{-2} I_{p+1}$. Comparing this with [5, Theorem 4.6.6] then proves the theorem. \square

2.2 Sectional curvatures of the linear regression model

Theorem 1 allows us to see that the linear regression parameter space Φ is a Riemannian manifold with all sectional curvatures equal to $-1/2n$. For if $\lambda > 0$ and (M, g) is a Riemannian manifold, where M is a smooth manifold and g is a metric tensor, then the sectional curvatures of (M, g) are λ^{-1} times the corresponding section curvatures of the Riemannian manifold $(M, \lambda g)$. (This is elementary to prove from the relevant definitions, but as a check that the correct power of λ here is -1 , apply this formula to the case when (M, g) is the unit 2-sphere: for then $(M, \lambda g)$ is isometric to the 2-sphere with radius $\sqrt{\lambda}$ and this has all sectional curvatures equal to λ^{-1} .) Combining this scaling result with Theorem 1 and the fact that the sectional curvatures of the upper half-space model are all equal to -1 then proves that Φ has all sectional curvatures equal to $-1/2n$.

2.3 The spherical normal model

The linear regression model can be viewed as a sub-model of the n -dimensional spherical normal model $y \sim N_n(\mu, \sigma^2 I_n)$ with unknown σ . On the other hand, the spherical normal model is the special case of the linear regression model where $p = n$ and $A = B = I_n$. Our finding from Section 2.2 that all linear regression models with n observations and p covariates have sectional curvatures of $-1/2n$ therefore contradicts Kass and Vos [4, §7.4.3] when $n \neq 1$, since they report that the sectional curvatures for the model $y \sim N_n(\beta, \sigma^2 I_n)$ are all $-1/2$ for all n . However, we will now show that this result in [4] cannot be correct.

Intuitively, when n is large, we would expect the n -dimensional spherical normal model (with a fixed number $N \neq 1$ of observations) to behave like the spherical normal model with known σ . But the σ -known model has a Euclidean geometry and hence sectional curvatures of 0, so the sectional curvatures for the the n -dimensional spherical normal model should approach 0 as $n \rightarrow \infty$. This is consistent with our result but not with that of [4].

A more careful argument can be given by interpreting the model for n independent and identically distributed univariate normal random variables $y_1, \dots, y_n \sim N(\mu, \sigma^2)$ as a sub-model of the n -dimensional spherical normal model. If we define $f(\mu, \sigma) \stackrel{\text{def}}{=} (\mu/\sqrt{2}, \dots, \mu/\sqrt{2}, \sigma)$ then f maps the μ, σ parameterisation of the former model into Kass and Vos' z parameterisation of the latter model (in a way that respects likelihood functions). The Jacobian matrix J of f is

$$J = \begin{bmatrix} \vec{1}/\sqrt{2} & \vec{0} \\ 0 & 1 \end{bmatrix}$$

where $\vec{1}$ and $\vec{0}$ are $n \times 1$ column matrices with all entries equal to 1 and 0, respectively. So by the change-of-variables formula (Lemma 6, below), if the formulae of [4, §7.4.3] were true then the Fisher information metric for n independent and identically distributed univariate normal random variables would be

$$J^T(2\sigma^{-2}I_{n+1})J = 2\sigma^{-2}J^TJ = 2\sigma^{-2} \begin{bmatrix} \vec{1}^T/\sqrt{2} & 0 \\ \vec{0}^T & 1 \end{bmatrix} \begin{bmatrix} \vec{1}/\sqrt{2} & \vec{0} \\ 0 & 1 \end{bmatrix} = \sigma^{-2} \begin{bmatrix} n & 0 \\ 0 & 2 \end{bmatrix},$$

which cannot be correct because the Fisher information matrix should scale linearly with the sample size.

In a similar way, we can see that the Fisher information matrix of [2, §II(i)] is also incorrect. This has been corrected in [3], though the sectional curvatures for the spherical normal model are not correct in either paper.

3 The distribution of the sufficient statistic

If y is a realization of a random variable Y then the sufficient statistic $x \stackrel{\text{def}}{=} T(y)$ is a realization of a different random variable $X = T(Y)$. It is a remarkable fact for exponential

families [1, p. 127], provable by a direct application of the smooth co-area formula, that the PDF $p_X(x|\theta)$ of X given θ is very similar to that of Y , namely

$$p_X(x|\theta) = \exp(\theta \cdot x) h_X(x) / Z(\theta) \quad (8)$$

where $h_X(x)$ is some function of x (which is not closely related to h_Y , in general). Therefore the PDFs for X given θ form a natural exponential family with the same natural parameter and the same partition function as the exponential family for Y .

Let \mathcal{X} be the set of all sufficient statistics, i.e., let \mathcal{X} be the image T as given in (2).

Lemma 2. *When $p < n$, \mathcal{X} is the solid paraboloid*

$$\mathcal{X} = \{x \in \mathbb{R}^{p+1} \mid x_{p+1} \geq x_1^2 + \dots + x_p^2\} \quad (9)$$

and when $p = n$, \mathcal{X} is the paraboloid $\{x \in \mathbb{R}^{p+1} \mid x_{p+1} = x_1^2 + \dots + x_p^2\}$.

Proof. Deferred to the Appendix. \square

We can now calculate the distribution of X given θ . In light of (8), this amounts to finding $h_X(x)$, though our proof will also establish (8) for linear regression.

Lemma 3. *The PDF $p_X(x|\theta)$ of X given θ is as in (8) where*

$$h_X(x) = c_h (x_{p+1} - x_1^2 - \dots - x_p^2)^{\frac{n-p-2}{2}}$$

and the constant $c_h = (2^{\frac{n}{2}} \pi^{p/2} \Gamma(\frac{n-p}{2}))^{-1}$, with Γ being the gamma function.

Proof. Deferred to the Appendix. \square

4 The SMML estimator for linear regression

In this section, we first recall the definition of the SMML estimator, which is a Bayesian estimator motivated by information-theoretic considerations. We then describe the expectation parameter space of the linear regression model and show that this can be naturally identified with the space \mathcal{X} of sufficient statistics. By Section 2, this gives \mathcal{X} a hyperbolic metric, and we finish by showing that an SMML estimator corresponds to a partition of \mathcal{X} into hyperbolic polytopes.

4.1 SMML estimators

The SMML estimator with m regions is defined as follows, where $m \geq 1$ is an integer [6, Chapter 3]. Suppose we are given a partition U_1, \dots, U_m of \mathcal{X} , parameters $\theta_1, \dots, \theta_m \in \Theta$ (the assertions) and real numbers $q_1, \dots, q_m \in \mathbb{R}$ (the coding probabilities for the assertions) so that $1 = q_1 + \dots + q_m$ and each $q_i > 0$. Let $\hat{\theta}$ and \hat{q} be the step functions given by $\hat{\theta}(x) \stackrel{\text{def}}{=} \theta_i$ and $\hat{q}(x) \stackrel{\text{def}}{=} q_i$ where i is the unique integer for which $x \in U_i$. If the data space \mathcal{X} is countable then we can use this structure to transmit any data point $x \in \mathcal{X}$ to an imaginary receiver by first transmitting the assertion $\hat{\theta}(x)$ using an optimal codebook constructed from the coding probabilities q_1, \dots, q_m , and second transmitting x using an optimal coding based on the assertion $\hat{\theta}(x)$. For linear regression, \mathcal{X} is not countable, so we simply truncate all data points to a finite but large number N of binary places and proceed as above [6, p. 167–168]. Then the (idealized) length of the assertion for x is $-\log \hat{q}(x)$ and the length of the detail is $-\log p(x|\hat{\theta}(x))$, so the average length of the message used to encode x is

$$I_1 = -\mathbb{E}[\log \hat{q}(X) + \log f(X|\hat{\theta}(X))] \quad (10)$$

plus the constant $N \log 2$ [6, p. 168]. Here, X is a random variable distributed according to the *marginal PDF*

$$r(x) \stackrel{\text{def}}{=} \int_{\Theta} \pi_{\Theta}(\theta) p_X(x|\theta) d\theta.$$

Definition 1. An SMML estimator with m regions is the function $\hat{\theta}(x)$ corresponding to any partition U_1, \dots, U_m , assertions $\theta_1, \dots, \theta_m$ and coding probabilities q_1, \dots, q_m which minimize I_1 .

Note that an SMML estimator with m regions might not exist or might not be unique in general, however we will often refer to ‘the’ SMML estimator when discussing this estimator informally.

Wallace [6, p. 156] gave conditions which the U_1, \dots, U_m , $\theta_1, \dots, \theta_m$ and q_1, \dots, q_m for an SMML estimator must satisfy. In the case of an exponential family with PDF of the general form (8), these are

$$U_i = \{x \in \mathcal{X} \mid \lambda_i(x) \leq \lambda_j(x) \text{ for all } j = 1, \dots, m\} \quad (11)$$

$$q_i = \int_{U_i} r(x) dx \quad (12)$$

$$\theta_i = f_{\Xi\Theta}^{-1} \left(\frac{1}{q_i} \int_{U_i} x r(x) dx \right) \quad (13)$$

where λ_i is the linear function of x given by $\lambda_i(x) = -\log q_i - x \cdot \theta_i + \log Z(\theta_i)$ and $f_{\Xi\Theta}$ is an invertible function which will be defined in Section 4.2, below.

Note that (11) shows that each U_i is a convex polytope (with respect to the affine structure on \mathcal{X} inherited from its ambient vector space). So U_1, \dots, U_m is a partition of \mathcal{X} into convex polytopes.

4.2 The expectation parameter space and its identification with the space of sufficient statistics

The expectation parameter ξ corresponding to the natural parameter θ is defined to be the expected value $\mathbb{E}[X|\theta]$ of X given θ , i.e., the expected value of a random variable with the PDF $p_X(x|\theta)$ given in Lemma 3. Let Ξ be the space of all expectation parameters and let $f_{\Xi\Theta} : \Theta \rightarrow \Xi$ be the map between the natural and expectation parameterisations, that is,

$$f_{\Xi\Theta}(\theta) \stackrel{\text{def}}{=} \int_{\mathcal{X}} x p_X(x|\theta) dx. \quad (14)$$

Since (14) expresses the expectation parameter $\xi = f_{\Xi\Theta}(\theta)$ corresponding to θ as a convex combination of elements of \mathcal{X} , it is clear that ξ lies in the same vector space as \mathcal{X} . In the case of linear regression when $p < n$, \mathcal{X} is convex by (9), so (14) further implies that $\xi \in \mathcal{X}$. So in our main case of interest,

$$\Xi \subseteq \mathcal{X}. \quad (15)$$

In fact, it is known that the expectation parameter space Ξ can be naturally identified with the interior of \mathcal{X} for many exponential families [1, Corollary 9.6]. We will sketch a proof of this fact, in the case of linear regression, after calculating the reparameterisation map $f_{\Xi\Theta}$.

By a standard result for exponential families (e.g. see [4, Theorem 2.2.1]), the partition function Z is infinitely differentiable, $f_{\Xi\Theta}$ can be calculated as

$$f_{\Xi\Theta}(\theta) = \nabla|_{\theta} \log Z \quad (16)$$

(where $\nabla|_{\theta} \log Z$ is the gradient of $\log Z$ evaluated at θ) and $f_{\Xi\Theta}$ is a diffeomorphism (i.e., an infinitely differentiable function with an infinitely differentiable inverse) from Θ to Ξ .

So from (4) and (16) we have

$$f_{\Xi\Theta}(\theta) = \frac{1}{-2\theta_{p+1}} \left(\theta_1, \dots, \theta_p, n + \frac{\theta_1^2 + \dots + \theta_p^2}{-2\theta_{p+1}} \right). \quad (17)$$

It follows easily from this and the defining property of Ξ (that Ξ is the image of $f_{\Xi\Theta}$) that

$$\Xi = \{\xi \in \mathbb{R}^{p+1} \mid \xi_{p+1} > \xi_1^2 + \dots + \xi_p^2\},$$

so comparing this to (9) and using (15) shows that Ξ is the interior of \mathcal{X} , i.e., up to a set with zero Lebesgue measure, there is a natural identification $\Xi = \mathcal{X}$ (when $p < n$). So since Ξ has a natural hyperbolic metric (by Section 2.1 and the fact that reparameterisation maps are isometries), this means that the interior of \mathcal{X} has one, too.

Table 1 gives the reparameterisation maps between the three parameterisations introduced so far.

4.3 Affine and hyperbolic lines in the expectation parameter space

We have just shown that the interior of the data space \mathcal{X} can be naturally identified with the expectation parameter space Ξ . We will now describe the relationship between the hyperbolic structure on Ξ (coming from the Fisher information metric) and the affine structure on Ξ (inherited from the vector space \mathbb{R}^{p+1} containing Ξ). We will show there is a natural function $H_{\Xi} : \Xi \rightarrow \Xi$ which maps affine lines in Ξ to hyperbolic lines in Ξ . Since the partition U_1, \dots, U_m corresponding to an SMML estimator consists of affine convex polytopes (by Section 4.1), this shows that $H_{\Xi}(U_1), \dots, H_{\Xi}(U_m)$ is essentially a partition of the hyperbolic space \mathcal{X} into hyperbolic convex polytopes.

Here, an *affine* plane P is the non-empty set, in $\Xi \subseteq \mathbb{R}^{p+1}$, of solutions to a set of possibly non-homogeneous linear equations. A *hyperbolic* plane Q is any subset of the interior of Ξ which contains the hyperbolic line (the image of a geodesic) through any two points of Q . Note that in this terminology, affine and hyperbolic lines are just 1-dimensional affine and hyperbolic planes (respectively).

Define $H_{\Xi} : \Xi \rightarrow \Xi$ to be $H_{\Xi} = f_{\Xi\Phi} \circ H_{\Phi} \circ f_{\Xi\Phi}^{-1}$ where $H_{\Phi} : \Phi \rightarrow \Phi$ is given by

$$H_{\Phi}(\phi) \stackrel{\text{def}}{=} (\phi_1, \dots, \phi_p, \phi_{p+1}/\sqrt{2}) \quad (18)$$

and $f_{\Xi\Phi} : \Phi \rightarrow \Xi$ is the reparameterisation map between Φ and Ξ , i.e.,

$$f_{\Xi\Phi}(\phi) = (\phi_1, \dots, \phi_p, \phi_1^2 + \dots + \phi_p^2 + \frac{\phi_{p+1}^2}{2}), \quad (19)$$

as can be calculated from the reparameterisation maps (2), (7) and (17) (see Table 1).

The map H_{Ξ} can be interpreted in terms of the hyperbolic geometry as follows. In the linear regression model, the point at infinity ∞ is a distinguished point on the sphere at infinity of the upper half-space Φ , and H_{Φ} translates each point $\phi \in \Phi$ away from ∞ along the geodesic through ∞ and ϕ by a distance $\log \sqrt{2}$. Since this description only depends on the distinguished point ∞ and notions from hyperbolic geometry, which are both preserved by $f_{\Xi\Phi}$, the same interpretation holds for H_{Ξ} .

Lemma 4. *P is an affine plane of Ξ if and only if $H_{\Xi}(P)$ is a hyperbolic plane of Ξ . In particular, H_{Ξ} maps affine lines to hyperbolic lines.*

Proof. The upper half-space model of hyperbolic $(p+1)$ -dimensional space coincides with Φ and the metrics on the two Riemannian manifolds are constant multiples of each other, so a hyperbolic plane of one is a hyperbolic plane of the other. But the p -dimensional hyperbolic planes of the upper half space model all have a known form [5], so the hyperbolic p -planes in Φ are of the form

$$Q = \{\phi \in \Phi \mid (\phi_1 - c_1)^2 + \dots + (\phi_{p+1} - c_{p+1})^2 = R^2\} \quad (20)$$

	Natural θ	Expectation ξ	Upper half space ϕ
$\theta =$	θ	$\frac{n}{V(\xi)}(\xi_1, \dots, \xi_p, -\frac{1}{2})$	$\frac{2n}{\phi_{p+1}^2}(\phi_1, \dots, \phi_p, -\frac{1}{2})$
$\xi =$	$\frac{1}{-2\theta_{p+1}}(\theta_1, \dots, \theta_p, n + \frac{\theta_1^2 + \dots + \theta_p^2}{-2\theta_{p+1}})$	ξ	$(\phi_1, \dots, \phi_p, \phi_1^2 + \dots + \phi_p^2 + \frac{\phi_{p+1}^2}{2})$
$\phi =$	$\frac{1}{-2\theta_{p+1}}(\theta_1, \dots, \theta_p, \sqrt{-4n\theta_{p+1}})$	$(\xi_1, \dots, \xi_p, \sqrt{2V(\xi)})$	ϕ
$H(\cdot)$	2θ	$(\xi_1, \dots, \xi_p, \xi_{p+1} - V(\xi)/2)$	$(\phi_1, \dots, \phi_p, \phi_{p+1}/\sqrt{2})$
$H^{-1}(\cdot)$	$\theta/2$	$(\xi_1, \dots, \xi_p, \xi_{p+1} + V(\xi))$	$(\phi_1, \dots, \phi_p, \sqrt{2}\phi_{p+1})$

Table 1: Maps between different parameterisations of the linear regression model, as well as some other useful quantities, where $V(\xi) = \xi_{p+1} - \xi_1^2 - \dots - \xi_p^2$.

or

$$Q = \{\phi \in \Phi \mid c_1\phi_1 + \dots + c_{p+1}\phi_{p+1} = d\} \quad (21)$$

for some $R > 0$, $d \in \mathbb{R}$ and $c \in \mathbb{R}^{p+1}$ with $c_{p+1} = 0$. And since $f_{\Xi\Phi}$ is an isometry, the hyperbolic p -planes in Ξ are all of the form $f_{\Xi\Phi}^{-1}(Q)$ for some hyperbolic plane Q in Φ .

Now, $P \subseteq \Xi$ is an affine p -plane if and only if $P \neq \emptyset$ and

$$P = \{\xi \in \Xi \mid L(\xi) = 0\}$$

for some (affine) linear function $L : \Xi \rightarrow \mathbb{R}$, say $L(\xi) = a_1\xi_1 + \dots + a_{p+1}\xi_{p+1} + b$. So

$$\begin{aligned} (L \circ f_{\Xi\Phi} \circ H_{\Phi}^{-1})(\phi) &= (L \circ f_{\Xi\Phi})(\phi_1, \dots, \phi_p, \sqrt{2}\phi_{p+1}) \text{ by (18)} \\ &= L(\phi_1, \dots, \phi_p, \phi_1^2 + \dots + \phi_p^2 + \phi_{p+1}^2) \text{ by (19)} \\ &= a_1\phi_1 + \dots + a_p\phi_p + a_{p+1}(\phi_1^2 + \dots + \phi_p^2 + \phi_{p+1}^2) + b \\ &= a_{p+1}((\phi_1 - c_1)^2 + \dots + (\phi_{p+1} - c_{p+1})^2 - R^2) \end{aligned} \quad (22)$$

if $a_{p+1} \neq 0$, where $c_{p+1} = 0$, $c_i = -a_i/2a_{p+1}$ for $i = 1, \dots, p$ and $R^2 = -b/a_{p+1} + c_1^2 + \dots + c_p^2$. Note that $R^2 > 0$ because $P \neq \emptyset$ so L has a zero in Ξ and hence $L \circ f_{\Xi\Phi} \circ H_{\Phi}$ must have a zero in Φ . Comparing (22) with (20) when $a_{p+1} \neq 0$, or comparing a similar expression with (21) when $a_{p+1} = 0$, shows that

$$\{\phi \in \Phi \mid (L \circ f_{\Xi\Phi} \circ H_{\Phi}^{-1})(\phi) = 0\} \text{ is a hyperbolic } p\text{-plane in } \Phi. \quad (23)$$

Now, if U and V are any two sets and $f : U \rightarrow V$ and $g : U \rightarrow \mathbb{R}$ are any functions with f injective (one-to-one) then

$$f(\{u \in U \mid g(u) = 0\}) = \{v \in V \mid g(f^{-1}(v)) = 0\}.$$

Applying this to the case $f = H_{\Phi} \circ f_{\Xi\Phi}^{-1}$, $g = L$, $U = \Xi$ and $V = \Phi$ gives

$$\begin{aligned} H_{\Xi}(P) &= H_{\Xi}(\{\xi \in \Xi \mid L(\xi) = 0\}) \\ &= (f_{\Xi\Phi} \circ H_{\Phi} \circ f_{\Xi\Phi}^{-1})(\{\xi \in \Xi \mid L(\xi) = 0\}) \\ &= f_{\Xi\Phi}(\{\phi \in \Phi \mid (L \circ f_{\Xi\Phi} \circ H_{\Phi}^{-1})(\phi) = 0\}) \\ &= f_{\Xi\Phi}(Q) \end{aligned}$$

where Q is a hyperbolic p -plane in Φ by (23). Therefore $H_{\Xi}(P) = f_{\Xi\Phi}(Q)$ is a hyperbolic p -plane in Ξ . Also, any hyperbolic p -plane arises in such a way, so this proves the lemma for p -dimensional affine and hyperbolic planes. So lastly note that an affine or hyperbolic plane of any dimension can be expressed as an intersection of p -dimensional planes, and that such intersections always give planes, so this proves the lemma. \square

5 The Jeffreys prior and the marginal distribution

In this section we will put the (improper) Jeffreys prior $\pi_{\Theta}(\theta)$ on θ and calculate the marginal distribution of X , i.e., the distribution of X not conditioned on θ . We choose the Jeffreys prior because it is natural, it makes few assumptions about the parameter values (i.e., it is uninformative) and it is tractable to work with. It also has a geometrical interpretation, so this choice preserves the symmetries of, and hence the close connections with, the underlying hyperbolic geometry.

5.1 The Jeffreys prior on the natural parameter space

From the definition (6) and the expression (3), it is easy to see that the Fisher information matrix g_{Θ} corresponding to the natural parameterisation of the linear regression model

(or any other exponential family [4]) is the Hessian of the log-partition function. So from (4),

$$g_{\Theta} = \frac{1}{-2\theta_{p+1}} \begin{bmatrix} I_p & -\theta_{p+1}^{-1}\theta_{[1:p]} \\ -\theta_{p+1}^{-1}\theta_{[1:p]}^T & -n\theta_{p+1}^{-1} + \theta_{p+1}^{-2}(\theta_1^2 + \dots + \theta_p^2) \end{bmatrix} \quad (24)$$

where $\theta_{[1:p]}$ is the $p \times 1$ column matrix with entries $\theta_1, \dots, \theta_p$. Recall that $\theta_{p+1} < 0$ so all entries of g_{Θ} are positive. The (improper) Jeffreys prior is defined to be $\pi_{\Theta}(\theta) \stackrel{\text{def}}{=} \sqrt{\det g_{\Theta}}$, so using (24) and expanding the determinant of g_{Θ} along its bottom row gives

$$\pi_{\Theta}(\theta) = \sqrt{n} 2^{-\frac{p+1}{2}} (-\theta_{p+1})^{-\frac{p+2}{2}}. \quad (25)$$

5.2 The marginal distribution

We can now calculate the marginal distribution of X (not conditioned on θ), whose PDF $r(x)$ is defined to be

$$r(x) \stackrel{\text{def}}{=} \int_{\Theta} \pi_{\Theta}(\theta) p_X(x|\theta) d\theta.$$

Lemma 5. *If $\pi_{\Theta}(\theta)$ is the Jeffreys prior then the marginal distribution of X is*

$$r(x) = c_r (x_{p+1} - x_1^2 - \dots - x_p^2)^{-\frac{p+2}{2}}$$

where $c_r = \sqrt{n} 2^{\frac{p-1}{2}} \Gamma\left(\frac{n}{2}\right) / \Gamma\left(\frac{n-p}{2}\right)$ and Γ is the gamma function.

Proof. Deferred to the Appendix. □

5.3 The marginal PDF is a multiple of the hyperbolic volume density

We will now show that the marginal PDF on \mathcal{X} corresponding to the Jeffreys prior is a constant multiple of the hyperbolic volume density (recall that the interior of \mathcal{X} has a natural hyperbolic metric by Section 4.2).

Recall that Θ and Ξ are the natural and expectation parameterisations of the linear regression model, that their Fisher information matrices are g_{Θ} and g_{Ξ} (respectively) and that the reparameterisation map $f_{\Xi\Theta} : \Theta \rightarrow \Xi$ between them is given by (17). Let $\pi_{\Xi}(\xi) = \sqrt{\det g_{\Xi}}$ be the volume density (i.e., the improper Jeffreys prior) on Ξ . Then a standard result for exponential families [4, Theorem 2.2.5] is that $g_{\Xi}(\xi) = g_{\Theta}^{-1}(f_{\Xi\Theta}^{-1}(\xi))$ where g_{Θ}^{-1} is the matrix inverse of g_{Θ} and $f_{\Xi\Theta}^{-1}$ is the inverse function of $f_{\Xi\Theta}$. Therefore $\pi_{\Xi}(\xi) = (\det g_{\Theta}^{-1}(f_{\Xi\Theta}^{-1}(\xi)))^{\frac{1}{2}} = (\pi_{\Theta}(f_{\Xi\Theta}^{-1}(\xi)))^{-1}$. It is easy to show from (17) that

$$f_{\Xi\Theta}^{-1}(\xi) = \frac{n}{\xi_{p+1} - \xi_1^2 - \dots - \xi_p^2} \left(\xi_1, \dots, \xi_p, -\frac{1}{2} \right)$$

so, from (25),

$$\begin{aligned} \pi_{\Xi}(\xi) &= n^{-\frac{1}{2}} 2^{\frac{p+1}{2}} \left(\frac{n}{2(\xi_{p+1} - \xi_1^2 - \dots - \xi_p^2)} \right)^{\frac{p+2}{2}} \\ &= n^{\frac{p+1}{2}} 2^{-\frac{1}{2}} (\xi_{p+1} - \xi_1^2 - \dots - \xi_p^2)^{-\frac{p+2}{2}} \\ &= \left(\frac{\Gamma\left(\frac{n-p}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \left(\frac{n}{2}\right)^{\frac{p}{2}} \right) r(\xi) \text{ by Lemma 3.} \end{aligned}$$

Up to a constant factor, the marginal probability $r(x)$ is therefore the hyperbolic volume density. Furthermore, it is not hard to see that the factor is approximately 1 when $p \ll n$ and p and n are even.

A Proofs of technical lemmas

We begin with a lemma which shows that Fisher information matrices behave well under reparameterisations, inclusions and submersions. In particular, this will show that the Fisher information matrices determine a well-defined metric on the underlying stochastic manifold (though it is not hard to prove this fact directly by giving a coordinate-free definition of the metric).

Let U and V be parameter spaces (of arbitrary dimensions) for two stochastic models and let $\ell_U : U \rightarrow \mathbb{R}$ and $\ell_V : V \rightarrow \mathbb{R}$ be the corresponding log-likelihood functions. If $f : U \rightarrow V$ is a function so that $\ell_U = \ell_V \circ f$ then we say that f maps U into V as a parameterised sub-model.

Lemma 6. *If $f : U \rightarrow V$ is a differentiable map which maps U into V as a parameterised sub-model then*

$$g_U = J^T g_V J$$

where g_U and g_V are the Fisher information matrices of the two parameterisations and J is the Jacobian matrix of f . In other words, g_U is the pull-back of g_V via f .

Proof. By definition, $g_U = \mathbb{E}[(\nabla \ell_U)(\nabla \ell_U)^T]$ and $g_V = \mathbb{E}[(\nabla \ell_V)(\nabla \ell_V)^T]$. By the chain rule, $\nabla \ell_U = J^T \nabla \ell_V$, where $\nabla \ell_U$ and $\nabla \ell_V$ are gradients of ℓ_U and ℓ_V . Therefore

$$g_U = \mathbb{E}[(\nabla \ell_U)(\nabla \ell_U)^T] = \mathbb{E}[J^T (\nabla \ell_V)(\nabla \ell_V)^T J] = J^T \mathbb{E}[(\nabla \ell_V)(\nabla \ell_V)^T] J = J^T g_V J,$$

as required. \square

We now give proofs for some technical lemmas.

Proof of Lemma 2. Recall that BB^T is the orthogonal projection onto $\text{col } A$, so $1 - BB^T$ is the orthogonal projection onto the space perpendicular to $\text{col } A$ and hence

$$\|y\|^2 = \|BB^T y\|^2 + \|(1 - BB^T)y\|^2$$

by Pythagoras' theorem. Since the columns of B form an orthonormal basis for $\text{col } A$, $\|BB^T y\|^2 = \|B^T y\|_{\mathbb{R}^p}^2$, where the second norm is the Euclidean norm on \mathbb{R}^p (and, as above, the norm without a subscript is the Euclidean norm on \mathbb{R}^n). So substituting $\|BB^T y\|^2 = \|B^T y\|_{\mathbb{R}^p}^2 = \|(x_1, \dots, x_p)\|_{\mathbb{R}^p}^2 = x_1^2 + \dots + x_p^2$ and $\|y\|^2 = x_{p+1}^2$ into the above formula we obtain

$$x_{p+1}^2 = x_1^2 + \dots + x_p^2 + \|(1 - BB^T)y\|^2. \quad (26)$$

Since $\|(1 - BB^T)y\|^2 \geq 0$, (26) implies $x_{p+1}^2 \geq x_1^2 + \dots + x_p^2$ and hence that the image of T lies in \mathcal{X} .

On the other hand, if $p < n$ then there exists a non-zero vector v perpendicular to $\text{col } A$, so given any $x \in \mathcal{X}$, if we define $y = Bx_{[1:p]} + tv$ where $x_{[1:p]}$ is the $p \times 1$ column matrix with entries x_1, \dots, x_p and $t = \sqrt{x_{p+1}^2 - x_1^2 - \dots - x_p^2}$ then $T(y) = x$, so the image of T also contains \mathcal{X} . Here, $T(y) = x$ follows by using $B^T v = 0$ and $B^T B = I_p$ to show that $B^T y = B^T (Bx_{[1:p]} + tv) = x_{[1:p]}$ so $\|y\|^2 = \|BB^T y\|^2 + \|(1 - BB^T)y\|^2 = \|Bx_{[1:p]}\|^2 + \|(1 - BB^T)y\|^2 = \|x_{[1:p]}\|_{\mathbb{R}^p}^2 + \|(1 - BB^T)y\|^2 = x_1^2 + \dots + x_p^2 + t^2\|v\|^2. \quad \square$

Proof of Lemma 3. Let $x = T(y)$ be the sufficient statistic and let $x_{[1:p]}$ be the $p \times 1$ column matrix whose entries are the first p sufficient statistics, so $x_{[1:p]} = B^T y$ by (2). Then since y given β and σ is normally distributed, so is $x_{[1:p]}$. Also, the expected value of $x_{[1:p]}$ is $B^T \mathbb{E}[y] = B^T A \beta$ and the variance-covariance matrix of $x_{[1:p]}$ is

$$B^T \text{Var}(y) B = B^T (\sigma^2 I_n) B = \sigma^2 I_p.$$

So $x_{[1:p]} \sim N_p(B^T A \beta, \sigma^2 I_p)$ and the PDF of $x_{[1:p]}$ given θ is

$$p(x_1, \dots, x_p | \theta) = (2\pi\sigma^2)^{-p/2} \exp\left(-\frac{\|x_{[1:p]} - B^T A \beta\|_{\mathbb{R}^p}^2}{2\sigma^2}\right) \quad (27)$$

where $\|\cdot\|_{\mathbb{R}^p}^2$ is the Euclidean norm on \mathbb{R}^p (and recall that the norm $\|\cdot\|^2$ without a subscript is the Euclidean norm on \mathbb{R}^n).

Now, from (26) and an equation immediately preceding it, we have

$$x_{p+1} = x_1^2 + \dots + x_p^2 + \|(1 - BB^T)y\|^2$$

and $x_1^2 + \dots + x_p^2 = \|BB^T y\|^2$. But y is a normal random variable and $BB^T y$ and $(1 - BB^T)y$ are uncorrelated, hence they are independent and so are their norms $x_1^2 + \dots + x_p^2$ and $\|(1 - BB^T)y\|^2$. Therefore

$$x_{p+1} = x_1^2 + \dots + x_p^2 + \sigma^2 Q$$

where Q is a chi-squared random variable with $n - p$ degrees of freedom which is independent of x_1, \dots, x_p . So x_{p+1} given x_1, \dots, x_p and θ is a deterministic linear function of Q , hence its PDF $p(x_{p+1} | x_1, \dots, x_p, \theta)$ can be calculated from the PDF of Q and the change of variables formula for PDFs as

$$\frac{1}{\sigma^2 2^{\frac{n-p}{2}} \Gamma\left(\frac{n-p}{2}\right)} \left(\frac{x_{p+1} - x_1^2 - \dots - x_p^2}{\sigma^2}\right)^{\frac{n-p-2}{2}} \exp\left(-\frac{x_{p+1} - x_1^2 - \dots - x_p^2}{2\sigma^2}\right). \quad (28)$$

Combining (27) and (28) then gives the PDF of X given θ :

$$\begin{aligned} p_X(x | \theta) &= p(x_{p+1} | x_1, \dots, x_p, \theta) p(x_1, \dots, x_p | \theta) \\ &= \sigma^{-n} \exp\left(\frac{x_{p+1} - x_1^2 - \dots - x_p^2 + \|x_{[1:p]} - B^T A \beta\|_{\mathbb{R}^p}^2}{-2\sigma^2}\right) \\ &\quad \times \left(2^{\frac{n-p}{2}} \pi^{p/2} \Gamma\left(\frac{n-p}{2}\right)\right)^{-1} (x_{p+1} - x_1^2 - \dots - x_p^2)^{\frac{n-p-2}{2}} \\ &= \sigma^{-n} \exp\left(\frac{x_{p+1} - 2x_{[1:p]} \cdot B^T A \beta + \|B^T A \beta\|_{\mathbb{R}^p}^2}{-2\sigma^2}\right) h_X(x) \\ &= \exp(\theta \cdot x) h_X(x) / Z(\theta) \end{aligned}$$

by (2) and (4). □

Proof of Lemma 5. From (25) and Lemma 3,

$$r(x) = \sqrt{n} 2^{-\frac{p+1}{2}} h_X(x) \int_{\Theta} (-\theta_{p+1})^{-\frac{p+2}{2}} \exp(\theta \cdot x) \frac{1}{Z(\theta)} d\theta.$$

But from (4),

$$\begin{aligned}
\exp(\theta \cdot x)/Z(\theta) &= (-2\theta_{p+1})^{\frac{n}{2}} e^{\theta_{p+1}x_{p+1}} \exp\left(\theta_1x_1 + \dots + \theta_px_p + \frac{\theta_1^2 + \dots + \theta_p^2}{4\theta_{p+1}}\right) \\
&= (-2\theta_{p+1})^{\frac{n}{2}} e^{\theta_{p+1}x_{p+1}} \exp\left(\frac{1}{4\theta_{p+1}} \sum_{i=1}^p [4\theta_{p+1}\theta_ix_i + \theta_i^2]\right) \\
&= (-2\theta_{p+1})^{\frac{n}{2}} e^{\theta_{p+1}x_{p+1}} \exp\left(\frac{1}{4\theta_{p+1}} \sum_{i=1}^p [(\theta_i + 2\theta_{p+1}x_i)^2 - 4\theta_{p+1}^2x_i^2]\right) \\
&= (-2\theta_{p+1})^{\frac{n}{2}} e^{\theta_{p+1}(x_{p+1}-x_1^2-\dots-x_p^2)} \exp\left(\frac{1}{4\theta_{p+1}} \sum_{i=1}^p (\theta_i + 2\theta_{p+1}x_i)^2\right) \\
&= (-2\theta_{p+1})^{\frac{n}{2}} e^{\theta_{p+1}(x_{p+1}-x_1^2-\dots-x_p^2)} \exp\left(\frac{1}{4\theta_{p+1}} \|\theta_{[1:p]} + 2\theta_{p+1}x_{[1:p]}\|_{\mathbb{R}^p}^2\right) \\
&= (-2\theta_{p+1})^{\frac{n}{2}} e^{\theta_{p+1}(x_{p+1}-x_1^2-\dots-x_p^2)} (-4\pi\theta_{p+1})^{\frac{p}{2}} f(\theta_{[1:p]}) \\
&= 2^{\frac{n}{2}+p} \pi^{\frac{p}{2}} (-\theta_{p+1})^{\frac{n+p}{2}} e^{\theta_{p+1}(x_{p+1}-x_1^2-\dots-x_p^2)} f(\theta_{[1:p]})
\end{aligned}$$

where $f(\theta_{[1:p]})$ is the PDF for a normal random variable $N_p(-2\theta_{p+1}x_{[1:p]}, -2\theta_{p+1}I_p)$ evaluated at $\theta_{[1:p]}$. Therefore

$$\begin{aligned}
r(x) &= \sqrt{n} 2^{\frac{n+p-1}{2}} \pi^{\frac{p}{2}} h_X(x) \int_{\Theta} (-\theta_{p+1})^{\frac{n-2}{2}} e^{\theta_{p+1}(x_{p+1}-x_1^2-\dots-x_p^2)} f(\theta_{[1:p]}) d\theta \\
&= \sqrt{n} 2^{\frac{n+p-1}{2}} \pi^{\frac{p}{2}} h_X(x) \int_{-\infty}^0 (-\theta_{p+1})^{\frac{n-2}{2}} e^{\theta_{p+1}(x_{p+1}-x_1^2-\dots-x_p^2)} d\theta_{p+1} \text{ by (5)} \\
&= \sqrt{n} 2^{\frac{n+p-1}{2}} \pi^{\frac{p}{2}} h_X(x) \int_0^\infty e^{-st} t^{\frac{n-2}{2}} dt
\end{aligned}$$

where $t = -\theta_{p+1}$ and $s = x_{p+1} - x_1^2 - \dots - x_p^2$. But the Laplace transform of $t^{\frac{n-2}{2}}$ is $\Gamma\left(\frac{n}{2}\right) s^{-\frac{n}{2}}$, so

$$\begin{aligned}
r(x) &= \sqrt{n} 2^{\frac{n+p-1}{2}} \pi^{\frac{p}{2}} h_X(x) \Gamma\left(\frac{n}{2}\right) (x_{p+1} - x_1^2 - \dots - x_p^2)^{-\frac{n}{2}} \\
&= \sqrt{n} 2^{\frac{n+p-1}{2}} \pi^{\frac{p}{2}} \left(2^{\frac{n}{2}} \pi^{p/2} \Gamma\left(\frac{n-p}{2}\right)\right)^{-1} \Gamma\left(\frac{n}{2}\right) (x_{p+1} - x_1^2 - \dots - x_p^2)^{\frac{-n+n-p-2}{2}} \\
&= c_r (x_{p+1} - x_1^2 - \dots - x_p^2)^{-\frac{p+2}{2}}.
\end{aligned}$$

where we have again used Lemma 3. \square

References

- [1] O. Barndorff-Nielsen. *Information and exponential families in statistical theory*. John Wiley & Sons, Chichester, 1978.
- [2] S. I. R. Costa, S. A. Santos, J. E. Strapasson. *Fisher information matrix and hyperbolic geometry*. In M. J. Dinneen (ed.) the Proc. of IEEE ISOC ITW2005 on Coding and Complexity, p. 34-36.
- [3] S. I. R. Costa, S. A. Santos, J. E. Strapasson. *Fisher information distance: a geometrical reading*. arXiv:1210.2354v3 [stat.ME].
- [4] R. E. Kass and P. W. Vos. *Geometrical Foundations of Asymptotic Inference*. John Wiley & Sons, New York, 1997.
- [5] J. G. Ratcliffe, *Foundations of hyperbolic manifolds*, Springer, New York, 1994.
- [6] C. S. Wallace. *Statistical and Inductive Inference by Minimum Message Length*. Springer, 2005.